PROTEIN STRUCTURE PREDICTION PROBLEM IN BIOINFORMATICS USING NATURE INSPIRED COMPUTING TECHNIQUES

A

Synopsis

Submitted in the partial fulfilment of the requirement of

Doctor of Philosophy

In

Computer Science

Submitted by

Niharika Chaudhary

Under the Supervision of

Dr. Sanjay Saini

Assistant Professor

Department of Physics & Computer Science

Faculty of Science

DEPARTMENT OF PHYSICS AND COMPUTER SCIENCE,
FACULTY OF SCIENCE
DAYALBAGH EDUCATIONAL INSTITUTE (DEEMED UNIVERSITY)
DAYALBAGH, AGRA (UP) – 282005

2020

1. Introduction

Proteins are the main basic components of our life. Proteins form the basis of structures such as skin, hair and tendons that are responsible for regulating biochemical reactions and transporting molecules (Lesk et al., 2019). The protein sequence is composed of 20 different amino acids that are aligned in linear chains via peptide bonds, i.e. the primary sequence of proteins. The classification techniques of popular protein sequences involve the extraction of particular characteristics of the sequences; these characteristics depend on the structural and functional properties of the amino acids (Gao et al., 2018). Proteins are initially linear chains of amino acids that fold into their three-dimensional structures under the influence of various chemical and physical factors, which determine their biological functions and properties (Wang et al., 2017). Given the importance of this problem for human life, NMR spectroscopy and X-ray crystallography are the most commonly used to determine the native structures of proteins (Moult J et al., 2016). With the rapid growth in the size of the protein database (PDB), as mentioned above the method fails. Although these methods are reliable, they are not always feasible, as they are expensive and time-consuming. Therefore, to predict the native structure of a protein, given its primary sequence, is an important and challenging task in computational biology.

The primary structure is a linear sequence of amino acids linked together by peptide bonds. Proteins fold due to hydrophobic effect, Vander Waals interactions, electrostatic forces and hydrogen bond (J. Setubal et al., 1997). Secondary structures are three-dimensional structures characterized by a repetitive pattern of union. The fundamental elements of the secondary structure of the proteins are the helices α, the sheets β, the coils and the turns (Wang et al., 2017). Proteins that include these secondary structures can be further folded into the tertiary structure forming a packet of secondary structures, curves and rings. The tertiary structure is the global three-dimensional structure (Chatterjee et al., 2019). Sometimes, multiple chains of proteins bind and form hydrogen bonds between them, giving rise to quaternary structures. Therefore, the prediction of the protein structure (PSP) is the process to determine the three-dimensional structure of a protein from its amino acid sequence (Zhou et al., 2014). Due to the fact that the biological function of the protein is determined by its structure,

the prediction of the protein structure (PSP) is today the most important and stimulating problem of bioinformatics.

1.1 Bioinformatics

It is a very promising research area in the progress of research and development in complex areas, such as medicine, biology, environment, public health, agriculture, drug design etc. (Mathur et al., 2018). It is a mixture of computer science and molecular biology. Most problems in bioinformatics are NPs of a hard nature, so researchers used soft computing and artificial intelligence techniques to solve these problems. Swarm Intelligence (SI) has recently emerged as a family of Nature Inspired Computing (NIC) algorithms, known above all for their ability to produce low-cost, fast and reasonably accurate solutions to complex research problems (Rajakumar et al., 2016), three main objectives of bioinformatics are:

- *To develop algorithms and mathematical models to probe the relationships between members of a large set of biological data.*
- *To analyse and interpret the heterogeneous type of data, including nucleotide and amino acid sequences, protein domains and protein structures.*
- *To implement tools that enables the efficient storage, retrieval and management of high-volume biological databases*.

1.2 Relevance of NIC Algorithms in Bioinformatics

Nature Inspired Techniques (NIC) is an emerging technology that aims at developing new computing techniques by getting ideas that observes how nature behaves in various situations to solve complex problems. Especially, algorithms of Swarm Intelligence (SI) are majorly stochastic search and optimization techniques, guided by the principles of collective behavior and the self-organization of insect swarms (Cui et al., 2014). They are efficient, adaptable and robust research methods that produce almost optimal solutions and have a lot of implicit parallelism. Different tasks in bioinformatics involve the optimization of different criteria (such as conformation search spaces, alignment score, protein folding, etc.); making the application of the tools is more obvious and

appropriate (Al Kawam et al., 2017). Bioinformatics problems rarely require the exact optimal solution; rather, they need solid, fast and near-optimal solutions, which are known to be efficiently produced by algorithms. These are more tolerable when performing SI algorithms than when performing deterministic algorithms.

1.3 Problems in the Field of Bioinformatics

1.3.1 Human Genome Project (HGP)

The (HGP) Human Genome Project was the international collaborative research program whose objective was to understand complete mapping of all human genes (Pevsner et al., 2015). All genes in a human together are known as "genome". The most important event to give a great impetus to the development of bioinformatics was the Human Genome Project (HGP), started in 1988. The draft of the genome was completed in June 2000 and the sequencing of the final genome was completed in April 2003, two years before the scheduled date (Al Kawam et al., 2017). A "genome" refers to the entire DNA in a cell, or more abstractly, to the typical genetic composition of the cells of the members of a particular species. There was a research initiative called the "Human Genome Project" which generated a "reference genome" sequencing lots of DNA from multiple people and combining it to build a model of human being's typical genetic sequence. It provides a DNA sequence for each of the chromosomes and identifies gene positions and other important features.

1.3.2 Motif Detecting Problem

The discovery of reasons is the process of discovering a significant model of nucleotides or amino acids that share two or more molecules (Mathur et al., 2018). The method of detecting multiple patterns in biological sequence is used to compare the genomes of several species. All multi-gene approaches will only find common elements of genes of a given organism. There are some open problems that should be considered: first, how to find some known motifs in the given gene sequences; second, how to find some unknown motifs in the upstream adjustment region and, finally, how to find an unknown motif regulated by a known transcription factor (Lalwani et al., 2015). For the reason detection problem, some features are:

- *The models are relatively short.*
- *The patterns are not completely aligned in all sequences. Therefore, the variation must be represented by a probability matrix and no single sequence can exactly define these schemes.*
- *The exact positions of the models may not be important, since the relationships between molecules are more important (Verma et al., 2011).*
- *Insertion and cancellation operations do not occur, as there are drastic effects on the formation of schemes for these operations.*

1.3.3 The Multiple Sequence Alignment Problems (MSA)

A multiple sequence alignment (MSA) is a sequence alignment of three or more biological sequences, generally proteins, DNA or RNA. In general, it is assumed that the input set of query sequences has an evolutionary relationship based on which they share a link and descend from a common ancestor (Rubio-Largo et al., 2016). Given two X and Y sequences, a pair alignment indicates the positions of each sequence that are considered functionally or evolutionarily correlated. From a family of N sequences, we need to discover the common patterns of this family. To assess the quality of an alignment, a popular option is to use the SP score method (sum of pairs) (Lipman DJ et al., 1989). The SP score basically adds the replacement scores of all possible combinations for pairs of sequence characters in a column of a multiple sequence alignment (Cui et al., 2014).

1.3.4 The Construction of Phylogenetic Trees

Phylogenetic or evolutionary trees are schematic binary trees that show evolutionary interrelationships between different species believed to have a common ancestor (Mathur et al., 2018).The leaves of such a tree represent the current species, while the internal nodes represent the hypothetical ancestors. Phylogenetic trees can be rooted or not rooted. A rootless tree simply represents phylogenesis but does not provide an evolutionary path. In the case of a rooted tree, one of the internal nodes is used as an external group and, in essence, becomes the common ancestor of all other external nodes (Feng et al., 1990).

In a phylogenetic tree, phylogenies are reconstructed based on the comparison between current objects. The term object is used to indicate the units for which the phylogeny must be reconstructed (Setubal J et al., 1999).

1.3.5 The RNA Secondary Structure Prediction

Ribonucleic acid (RNA) is a nucleic acid polymer (such as DNA) consisting of nucleotide monomers. RNA serves as a model for the translation of genes into proteins, the transfer of amino acids into the ribosome to form proteins and even transcription into proteins.

Like the secondary structure of proteins, the secondary structure of RNA can conveniently be seen as an intermediate phase in the formation of a three-dimensional structure (Indumathy et al., 2012). The secondary structure of RNA is mainly composed of double-stranded RNA regions formed by folding the single-stranded molecule onto itself. To produce such double-stranded regions, a sequence downstream of the bases in the RNA must be complementary to another upstream sequence so that the coupling of the Watson-Crick bases can occur between the complementary GC and AU nucleotides (Ali et al., 2016). Among the various recognizable "domains" of the secondary structure, three knowns are the fork rings, the bumps and the inner rings. The secondary structure of RNA molecules can be computationally predicted by calculating the minimum free energy structure (MFE) for all the different combinations of bonds and hydrogen domains.

1.3.6 The 2D HP Protein Folding Problem

The hydrophobic-polar model (HP model) of the protein structure was first proposed by (Lau et al., 1989). It is motivated by a series of well-known facts on the fundamental role of hydrophobic and polar amino acids for the protein structure (Richards et al., 1977). Hydrophobic interaction is the driving force of protein folding and the hydrophobicity of amino acids is the main force for the development of a native conformation of small globular proteins (Li et al., 2016). Each of the twenty commonly found amino acids that are the basic components of all natural proteins can be classified as hydrophobic (H) or polar (P). Based on this classification, in the HP model, the primary amino acid sequence of a protein (which can be represented as a string on a twenty-letter alphabet) is subtracted in a sequence of hydrophobic and polar residues, that is, components of amino acids (Li et al.,2014).

1.3.7 Fragment Assembly Problem (FAP)

The problem of fragment assembly (FAP) concerns DNA sequencing. Currently, DNA strands of over 500 base pairs cannot be sequenced very precisely (Al Kawam et al., 2017). As a result, to sequence larger strands of DNA, they are first broken down into smaller pieces. The FAP is therefore to reconstruct the sequence of the original molecule from the sequences of smaller fragments. The FAP is basically a permutation problem, similar in spirit to the TSP, but with some important differences (circular paths, noise and special relationships between entities) (Meksangsouy P et al., 2003). It is also complete NP in nature. The authors studied two types of assembly problems: simple configuration problems and multiple configuration problems. The results of the simulation indicate that in the problems of a single contig, the algorithm of the system of colonial ants overcomes the nearest near heuristic algorithm when we consider more contig problems (Indumathy et al., 2012).

2. Problem Description

2.1 Protein Structure Prediction Problem

Proteins are macromolecules composed of 20 amino acids that are arranged in linear chains via peptide bonds, i.e. the primary sequence of proteins. These amino acids are the building blocks of all-natural proteins that are classified into (a) Hydrophobic (H) and (b) Polar (P). The prediction of the 3D structure of the protein from its primary sequence is established as a problem of protein structure prediction (PSP). PSP is one of the research areas open in structural bioinformatics. The unique 3D structure of the protein is of great importance, as it determines the biological functions and the properties of a protein (Lesk et al., 2019). So, understanding the 3D structure from its 1D amino acid sequence is the unsolved problem. There are several factors that make PSP difficult. These include a large conformation search spaces since the factor leading to protein folding is not well known. There are several approach problems, including the modeling of sequence homology, the Ab-initio approach and the Threading approach. The Ab-initio methods predict the 3D structure of proteins given their primary sequences without relying on protein databases (Wang et al., 2017).

The basic strategy is to find the best possible structure based on a chosen energy function. The main challenge of these approaches is to look for the most stable structure on a large search space and the most stable structure is the one with the lowest possible energy. Here we describe the prediction of the protein structure using the Ab-initio method, which is based on the global optimization of the potential energy function so holds the main focus in PSP problem (Dubey et al., 2016).

2.1.1 Protein Folding

During the synthesis, protein folds into unique 3D structures and this process is called as protein folding. To obtain the secondary structure (2D) of protein the primary sequence of a protein undergoes the process of folding and further they fold and obtain the tertiary structure (3D). Polypeptide chains are transformed into compact structures that perform biological functions, include control and regulation of essential chemical processes for the living organisms (Li, D et al., 2016). Under physiological conditions, the most stable three-dimensional structure is called the native conformation and allows a protein to perform its function, which, in turn, is a function of its primary structure. However, failure to fold into the unique three-dimensional structure usually leads to proteins with different properties to become inactive. In the worst case, such misfolding of proteins can be harmful to the organism (Khoury et al., 2014). For example, several diseases such as Alzheimer's disease, cystic fibrosis and some types of cancer are the results from the accumulation of misfolding of proteins. Therefore it is so important to study how proteins fold. There exist two protein folding problems.

- *To predict the protein structure (conformation) from sequence (primary structure)*
- *To predict protein folding pathways, this determines the folding sequence of proteins which lead from the primary structure of a protein (linear sequence of amino acids) to its native structure (Chatterjee et al., 2019).*

2.1.2 2D and 3D Hydrophobic-polar (HP) Model

Hydrophobic-polar (HP) is an NP-complete free energy model which is motivated by a series of well-known facts on the fundamental role of hydrophobic and polar amino acids for the structure of

proteins (Parpinelli et al., 2014). First proposed by (Dill et al., 1985), as a highly simplified model for predicting how protein folds in huge search space. In the HP model, the amino acid sequence is subtracted in a sequence of hydrophobic (H) and hydrophilic (P) residues. The protein conformations of this sequence are limited to the self-avoiding routes in a network; for our purpose, the HP 2D model considered here is a 2D square network and the HP 3D model consists of a 3D cubic network (Zhou et al., 2013). Based on the above biological motivation, the energy of a conformation is defined as a number of topological contacts between hydrophobic amino acids that are not close in the given sequence.

Hydrophobic interaction is the driving force of protein folding and the hydrophobia of amino acids is the main force for the development of a native conformation of small globular proteins (Gao et al., 2018). The native structures of many proteins are compact and have well-packed nuclei that are highly enriched in hydrophobic residues, as well as a minimum of non-polar surfaces exposed to solvents.

2.2 Computational Approaches to PSP

2.2.1 Homology Modeling: this is the simplest and most reliable approach. It is observed that proteins with the same sequences tend to bend in the same structures form the basis of this method. It can be seen that even proteins with a 25% sequence identity bend in the same structures. This approach does not work for remote counterparts (<25% peer identity). The method for homological modeling can be briefly defined as: given a sequence of Q queries and a database of sequences of known protein structures, find a protein P such that P has the same sequence as Q and returns the structure of P as an approximation to structure Q '. Following are the main steps in homology modeling (Hassanien et al., 2013):

- *Find known structures related to the sequence of queries whose structure must be modeled.*
- *Align the query sequence to the models.*
- *Build variable side chains and main chains.*
- *Refinement of the model, evaluation of the built model and choice of the most native conformations.*

2.2.2 Threading Approach: threading is an approach to folding recognition. It is used for sequences with sequence identity ≤ 30%. In this approach, given a sequence and the set of folds available in the Protein Data Bank (PDB), the goal is to see if the sequence can accept one of the folds of the known structure (Moult et al., 2016). This method is applied when the sequence has little or no similarity of the primary sequence with any sequence in a known structure or the structure library model represents the true fold of the sequence. It is a process to find similar folds and similar sequences. Initially, the method attempts to discover the structure of the known protein and find the fold most compatible with the sequence of queries whose structure must be predicted from a library. An optional match between the query sequence and a structural model will be determined by alignment algorithms and will evaluate the quality of a fold using an objective function. Higher scoring alignments provide possible models. The sub processing approach exceeds the comparative modeling for sequence similarities less than 25% (Rahman et al., 2005).

2.2.3 Ab-initio Approach: since the modeling and threading of homology requires the knowledge of known structures, the prediction of the Ab-initio structure has no limits like the previous approaches. Start with the estimate that the actual structure of a protein is at the minimum of the global free energy (Parpinelli et al., 2014). The Ab-initio or denovo method is a modeling approach without a model. It predicts the three-dimensional native conformation only from its primary sequence based on the hydrophobic and hydrophilic properties of the amino acids (Samudrala et al., 1999). The concept of the Ab-initio method is based on the thermodynamic hypothesis of Anfinsens according to which the native structure corresponds to the global minimum of a free energy function. All possible conformations must be evaluated and the structure with minimum energy is treated as a native structure (Dubey et al., 2016). Therefore, the representation of the conformation of a protein is very important in the Ab-initio approach. Each conformation is accessible via an energy function and, finally, a research technique is required to determine the native confirmation of the set of all possible confirmations.

3. Solution Techniques

There exist some solution techniques widely used to solve various bioinformatics problems.

3.1 Hidden Markov Model (HMM)

HMM was first introduced in the 1960s (Baum et al., 1966). Its applicability to bioinformatics problems became evident in the late 1990s (Krogh et al., 1998). It is a statistical model that can be used to describe the evolution of observable events that depend on internal factors, which are not directly observable. We call the observed event a "symbol" and the invisible factor at the base of observation a "state". HMM is also called a doubly incorporated stochastic process as it consists of two stochastic processes (a) an invisible process of hidden states (b) and a visible process of observable symbols. In a Markov chain, the hidden states and the probability distribution of the observed symbol depends on the underlying state.

The first probabilistic method that is considered the starting point of the prediction of the secondary structure is the "Chou-Fasman Method" (Chou-Fasman et al., 2014). The first attempt to use HMM for prediction of secondary protein structure was due to (Kiyoshi Asai et al., 1993). In his approach, four sub-models are trained separately, representing the four secondary structures: helix, sheet, turn and coil, PSP based on profile HMM and QPSO (Long et al., 2014).

3.2 Support Vector Machine (SVM)

SVM is a new machine learning method that has developed rapidly and has been widely used in many types of pattern recognition problems. The basic method of SVM is to transform the samples into a high-dimension Hilbert space and look for a separation plane in this space (Li, D et al., 2016). The separation hyperplane, called the optimal separation plateau (OSH), is chosen in such a way as to maximize its distance from the nearest training samples. SVM has been successfully applied to many fields of pattern recognition, including object recognition, identification of 17 speakers, 18 and text categorization.19 SVM generally outdoes other machine learning technologies, including neural networks and K-Neighbor classifiers (Guo et al., 2004). In recent years, SVM has been used in

bioinformatics, including the classification of the gene expression profile, the detection of remote protein homologies and the recognition of sites for the initiation of translation (Wang et al., 2017). A single SVM layer is used to analyse the secondary protein structure with excellent prediction results.

3.3 Combinatorial Optimization Technique

These are the problems in which a series of feasible solutions is discrete or can be reduced to make it discrete, so as to obtain the best optimal solution that minimizes the given function. Branch and Bound or Branch and Cut are some techniques to solve these problems, however they are not big and complicated problems. As the complexity of the problem increases, the cost of the algorithm increases considerably and the optimization of the solution is reduced, which makes the solution not feasible. Other methods to optimize these problems include the determination of a non-optimal solution at an economic time using heuristic and meta-heuristic strategies (Rajakumar et al., 2016).

3.4 Multi-objective Optimization Technique

The multi-objective is the procedure to simultaneously optimize two or more conflicting objectives that is subject to certain restrictions. It can be found in almost all practical conditions, where the solution received minimizes some goals and maximizes other goals at the same time (Cui et al., 2014). The single solution cannot simultaneously limit and maximize every goal of a well-formed multi-objective problem. Therefore, the solution must be such as to optimize each objective to a particular extent. Finding such solutions and quantifying them by evaluating others is the main goal in solving a multi-objective optimization problem (Rubio-Largo et al., 2016). The two preferred approaches to multi-objective optimization strategies are: 1. Scalar methods (classical approach based on preferences) 2. Pareto optimal method (ideal approach)

The scalar approach also called the choice-based method solves the problem by constructing an aggregate objective function which is the aggregation of all objectives (Tan et al., 2006). Here, the weight of a target is proportional to the vector of desire assigned to that goal. This method converts the problem of multiple objectives into a single-objective problem by resizing a target vector in a single composite objective function.

The second approach called the Pareto Optimal technique, which determines an uncontrolled series of solutions. When switching from one Pareto solution to another, there is a certain loss or advantage in one or more goals, so that you can get some advantage or loss in another. Pareto optimal solution sets are generally implemented in unique solutions because of their practical applications in real life problems. The size of Pareto sets generally increases when the number of goals increases. The relative desire vector (based on preferences) is highly subjective, so it is not so simple and easy to calculate, while Pareto's optimal technique is more realistic and less subjective.

3.5 Meta-heuristic Approaches

The meta-heuristic is the general algorithmic structure designed to solve complex optimization problems (Osman et al., 1996). For some decades, meta-heuristics has emerged as a successful alternative to classical heuristics to solve complex optimization problems. A meta-heuristic method tries to find an almost optimal solution, instead of trying specifically to find the exact optimal solution; generally, it does not have rigorous convergence tests with the optimal and computationally faster solution than the exhaustive search (Rajakumar et al., 2016). This class of algorithms has two main classifications which include: simulated annealing, tabu search (Glover et al., 1998), local search (Aarts et al., 2003); it lies in the Single Solution Approaches to the unique solution of meta-heuristic class. Another is the population-based approach, which includes evolutionary algorithms Genetic Algorithm (GA), Neural Network (NN) and Swarm Intelligence techniques, so various meta-heuristic methods were applied to bioinformatics applications (Kenneth et al., 2006) and (Ali et al., 2016).

3.5.1 Simulated Annealing (SA)

It is a relaxation approach for optimization problems, which identifies a precise approximation to global values (minimum or maximum) for the given function in the search space (Laarhoven et al., 1987). The implementation of meta-heuristics has become a successful algorithm, which provides good solutions (Osman et al., 1996); however, it did not compete with the tabu search algorithms proposed at that time. Tabu Search (TS) explores the space of the solution moving at each iteration of a solution ‘s’ to the best solution in a subset of its neighborhood N(s). The tabu search techniques have been widely proposed in the last 2-3 decades. The Local Search (LS) development heuristic begins

with an initial solution, explores the neighborhood to determine a better technique for updating the existing solution and repeats until finding better solutions in the neighbourhood (Aarts et al., 2003).

3.5.2 Neural Networks (NN)

Neural Networks (NN) are computational models consisting of small units interconnected with each other through some correlated weights, such as neurons within the brain (Ali et al., 2016). The artificial neural network (ANN) is a mathematical model based on biological NN. A self-organized map (SOM) is a form of ANN that is trained using unsupervised learning to produce low-size (usually 2-D) training samples, known as a map. Hopfield's pioneering work (Hopfield et al., 1986) for setting TSP offers a start to solve combinatorial problems. (Qian et al., 1988) introduced one of the first neural networks to predict the secondary structure of protein. Chandonia also used the standard amino acid sequence as input for their neural network, and used the output of this prediction along with other information to predict the tertiary structure (Chandonia et al., 1995). Subsequently, HeiDelber (PHD) introduced the prediction methodology of neural profile networks, which is considered the backbone of all current methods (Rost et al., 1996).

3.5.3 Evolutionary Approaches

Evolutionary approaches are a part of evolutionary calculations, based entirely on biological evolution and using iterative processes to determine the optimal solution, Genetic algorithms (GA) inspired by genetic processes in models of nature, chromosome, selection, mutation and crossed concepts (Jong et al., 2006). Genes encode proteins that in turn can regulate the expressions of other genes. If we compare, we can see that although caterpillars and butterflies have identical genomes, the difference in the expression of genes in the DNA leads to their obvious physical difference. These events motivate the modeling of gene expression networks (Verma et al., 2011). The evolutionary algorithm copies the natural evolutionary principles to guide your research towards an optimal solution. EA uses a population of solutions instead of a single solution for iteration and the final results are also the population of the solution, i.e. the optimal Pareto set. A solution search engine can select any solution from the optimal Pareto set thus obtained. EA usually solves optimization problems efficiently. GA

provides a metaphor for processes; therefore, applying the genetic algorithm for the prediction problem of the protein structure, we can predict from the metaphor the territory and the quaternary changes in the amino acid sequence (Rout et al., 2013).

3.5.4 Swarm Intelligence (SI) Techniques

Nature-inspired computing (NIC) is an emerging technology, inspired by the behavior of nature in numerous situations, to solve complex problems. Swarm Intelligence is a sub-area of the NIC and is based on studying the cooperative behavior of simple individuals (e.g. ants, flocks of birds and bees, etc.) in various decentralized systems (Das et al., 2008). The population includes simple individuals; they can usually solve complex tasks by interacting locally between individuals and their environments. Therefore, SI is a promising stream to develop powerful solutions for optimization problems.

Ant Colony Optimization (ACO): is a meta-heuristic approach stimulated by an analogy, which uses pheromone traces for shorter paths when looking for food (Dorigo et al., 1999). A moving ant that marks the path with a trace of some essence called pheromone on the ground. A single ant passes at random and when it detects a trail of established pheromones, it follows that path and reinforces the trail with its pheromone, which attracts more ants. Therefore, the method is characterized by a fine cycle, in which the probability of an ant choosing a path increases with the number of ants moved on that path (Das et al., 2008). Therefore, ACO methods are useful in problems that need to find paths to goals. The Ant Colony (ACO) algorithm based on the HP 2D model which first produces a random pheromone matrix, so some ants are considered as conformations was proposed by Shmygelska (A. Shmygelska et al., 2003).

Particle Swarm Optimization (PSO): is a population-based method proposed by Kennedy and Eberhart and simulated through the behavior of flocks of birds and fish (Kennedy et al., 1995). The PSO optimizes the problem by iteratively improving a population of candidate solutions called particles, moving within the vast research area; adjusting its position and speed based on nearby particles. The movement of each particle is influenced by its best known local position (pbest) and

guided through the most known global position (gbest) and by the speed of its neighbors in the search space. In this way, during the upgrade to get better positions in the neighborhood, the swarm moves towards the best optimal solutions. The PSO has been successfully implemented in many studies and application sectors of many bioinformatics problems and has achieved better results faster and less expensive than the various strategies (Chen et al., 2011). The real value PSO algorithm by (Liu et al., 2005) and (Zhu et al., 2009) was successfully applied in the protein structure prediction (PSP) problem with different models external to the network, such as the HP model with ECEPP power supply function.

Bees Algorithm: The bees algorithm was introduced by Karaboga and later Karaboga and Akay changed it with a new meta-heuristic approach (Karaboga et al., 2005) and (Karaboga et al., 2011). This algorithm is inspired by the intelligent behavior of bees that try to find food (nectar). The colony of bees is formed by a sovereign (queen) with numerous male drones and a wide variety of workers (non-reproductive). The queen's trial is to mate with drones and start new colonies. Used bees assemble the honeycomb. The duty of young bees is to feed the queen and male drones, clean and protect the entire colony and accumulate more food (Li et al., 2014). The collective interaction behavior of specialized individuals, the assigned simultaneous responsibilities and the self-organization lead to the colony as a system of prepared teamwork. The colony of artificial bees (ABC) is an iterative algorithm, in which each agent (ape) is assigned to a solution (food source) that is randomly generated. Thus, in each iteration, each agent uses the neighborhood operator to find a new solution. The attitudinal function (amount of nectar) of the newly determined solution is evaluated by iteration; if it is at its maximum, replace the previous solution with the most recent one. The use of Artificial Bee Colony (ABC) algorithm to solve the PSP problem was introduced by (Bai Li et al., 2014). They used the internal ABC-based feedback strategy, which turned out to be an effective approach to exploration over exploitation.

3.6 Emerging NIC Techniques

Bacteria Foraging Optimization Algorithm (BFOA)

The Bacterial Foraging Algorithm was proposed by Passino, K. M. (Passino et al., 2002). Bacteria search for the nutrient in a manner to maximize the energy that is obtained per unit time. The Bacterial Foraging Algorithm is based on the observations of E. coli bacteria in nature that has Diameter: 1μm and Length: 2μm. The E. coli bacteria eat food to persist. During its foraging, locomotion is achieved by a set of tensile flagella. Flagella are "decision making," sensors, & exhibits closed-loop behavior. Flagellum is Counter clockwise: Swim and Clockwise: Tumble. In a harmful place flagellum tumbles frequently to find a nutrient gradient. Each tumble or movement of a bacterium is known as a chemotaxic step. BFOA mimics this chemotactic movement of bacteria in the problem space.

Intelligent Water Drops (IWD)

Shah-Hosseini introduced Intelligent Water Drops (IWD) as an algorithm, inspired by the natural water drops flowing in rivers and the way they deviate to find an optimal path to their destination (Shah-Hosseini et al., 2007). IWD have two essential properties:

- *Velocity (certain initial velocity): less soil more velocity gains in IWD and vice-versa.*
- *Soil (initially zero): the soil is added to an IWD from one location to another*

In the IWD algorithm, artificial water drops change their environment to select the best path with least soil on their links, i.e. also so that the next location to move paths should have less soil. Therefore, the solutions are formulated incrementally using the IWD algorithm. Shah-Hosseini tested a modified IWD algorithm in TSP and received almost optimal solutions. Rayapudi has also applied the IWD algorithm to solve an economic load dispatch problem (Rayapudi et al., 2011).

Gravitational Search Algorithm (GSA)

Rashedia introduced the idea of GSA (Rashedia et al., 2009). Gravitational search algorithm (GSA) could be considered as an isolated system of masses. It is like a small artificial world of masses obeying the Newtonian laws of gravitation and motion. Searcher agents are the collection of masses.

Using the gravitational force, every mass in the system can see the situation of other masses. All these objects attract each other by a gravitational force, and this force causes a movement of all objects globally towards the objects with heavier masses. The heavy masses correspond to good solutions of the problem. The position of the agent corresponds to a solution of the problem, and its mass is determined using a fitness function. By lapse of time, masses are attracted by the heaviest mass. This mass would present an optimum solution in the search space.

River Formation Dynamics (RFD)

River formation dynamics (RFD) is a heuristic method similar to ant colony optimization (ACO). In fact, RFD can be seen as a gradient version of ACO. RFD is based on copying how water forms rivers by eroding the ground and depositing sediments (Rabanal et al., 2007). As water transforms the environment, altitudes of places are dynamically modified, and decreasing gradients are constructed. The gradients are followed by subsequent drops to create new gradients, reinforcing the best ones. Good solutions are given in the form of decreasing altitudes. This method has been applied to solve different NP-complete problems of finding a minimum distance tree and finding a minimum spanning tree in a variable-cost graph.

Artificial Immune Systems (AIS)

Artificial immune systems (AIS) concerns the usage of abstract structure and function of the immune system to computational systems. It investigates the application of these systems towards solving computational problems from mathematics, engineering, and information technology. Artificial immune system is a sub-field of Bio-inspired computing, and natural computation, with interests in Machine Learning and belonging to the broader field of Artificial Intelligence.

Stochastic Diffusion Search (SDS)

Stochastic diffusion search (SDS) is an agent-based probabilistic global search and optimization technique. Best suited to problems where the objective function can be decomposed into multiple independent partial-functions. Each agent maintains a hypothesis which is iteratively tested by

evaluating a randomly selected partial objective function parameterised by the agent's current hypothesis. Information on hypotheses is diffused across the population via inter-agent communication (Das et al., 2008). SDS agents communicate hypotheses via a one-to-one communication strategy. A positive feedback mechanism ensures that, over time, a population of agents stabilizes around the global-best solution.

3.7 Parallel Computation

In parallel computing, large and complex problems are divided into many small problems to reduce calculation times by solving them simultaneously. Usually, an algorithm is written and executed as a stream of serial instructions that are executed one by one in a computer system, which takes a long time. Here, parallel computing techniques are essential, allowing multiple processes to run simultaneously, since the problem is broken down into smaller processes and distributed to numerous processors. Parallel computer systems can be labeled as multi-core and multiprocessor computer systems that can perform more than one process in a single system, while clusters, grids and MPPs use exceptional computer systems for the same calculation. Parallel computing programs are more difficult to write than sequential programs because simultaneous computing introduces new software errors, such as the most common competition conditions (Jozefowiez et al., 2002). Furthermore, synchronization and verbal exchange between secondary activities represent one of the most important obstacles to a good parallel program.

4. Literature Review

Lamiche chaabane et al. discussed the application of artificial intelligence (AI) techniques as PSO and simulated annealing (SA) to solve the problem of multiple sequence alignment (MSA) (Lamiche chaabane et al., 2018). **Rubio-Largo et al.** proposed the Hybrid multi-objective artificial bee colony for MSA (Rubio-Largo et al., 2016). A new hybrid model called the PSOSA approach, the PSO is exploited in global research, but easily trapped in the local optimum and can lead to premature convergence. **Lalwani et al.** applied a novel two-level PSO approach for efficient MSA (Lalwani et al., 2015). SA is incorporated as a local improvement approach to overcome the optimal local problem

and intensify research in local regions to improve the quality of the solution. **Riaz T et al.** presented a tabu search algorithm to align multiple sequences and also developed a hybrid approach based on dynamic programming (DP) and the PSO algorithm to overcome computational complexity and optimal local problems for MSA (Riaz T et al., 2004). **Verma et al.** also applied PSO and DSAPSO to solve DNA sequence assembly problem **(**Verma et al., 2011).

Meksangsouy et al. attempted to solve the problem of reorganizing DNA fragments with the ant colony system (Meksangsouy, et. al., 2003). **Indumathy et al.** applied NIC algorithms to solve DNA fragment assembly problem. The authors studied two types of assembly problems: simple configuration problems and multiple configuration problems. The simulation results indicate that the ant colony algorithm exceeds the closest heuristic algorithm when considering multiple configuration problems (Indumathy et al., 2012). **Wetcharaporn et al.** present the use of a system of colonies of combined ants (SCA) and the nearest heuristic algorithm (NNH) in the assembly of DNA fragments (Wetcharaporn et al., 2006). **Angeleri et al.** describes an alternative approach to the problem of fragment assembly (Angeleri et al., 1999). **Pfeiffenberger et al.** predicted improved protein conformations with a temporal deep recurrent neural network. The key idea is to form a Recurrent Neural Network (RNN) to trace a sequence of bases that constitute a given fragment and assign to the same group all the sequences that are well traced by this network (Pfeiffenberger et al., 2018).

Chen et al. have made a serious attempt to solve the classic MSA problem using a partition approach together with the ant colonization optimization algorithm (ACO) and a genetic algorithm (Chen et al., 2006). **Rasmussen et al.** focused on a new PSO-based training method for Hidden Markov Models (HMM) to solve the MSA problem (Rasmussen et al., 2003). The authors showed that a combination of PSO and evolutionary algorithms can generate better alignments of the protein sequence compared to more traditional HMM training methods, such as Baum-Welch and simulated annealing (Stolcke et al., 1993) (Hamam et al., 1996). The genetic algorithm is one of the important and successful approaches in the MSA. An improved GA method was proposed such that, a small multiple initialization strategy (MSPIS) and a hybrid GA based on a one-point crossover scheme (HOPCS), which can look for solution space in a certain way very efficiently (Zhang et al., 2004). The

experimental results show that this improved approach can achieve a better result than the traditional GA approach in aligning the problem of multiple protein sequences. The widespread logic allows the tolerance of inaccuracies or errors in the correspondence of the sequence.

Nasser et al. proposes a fuzzy logic for an approximate adaptation of the sub-sequences (Nasser et al., 2007). A fuzzy set based adaptive neighborhood search optimization heuristic for protein structure prediction was proposed by (Blanco et al., 2002). A fuzzy generalization of contact map was presented in (Pelta et al., 2005).

Ayan Datta et al. attempted a new approach to predict the tertiary structure of proteins using the artificial neural network (ANN) and the particle swarm optimization technique (PSO).The main objective lies in the well-known Ab-initio approach to the global minimization of the energy function. Predict the native structure of the protein by finding the dihedral angles of the main chain; through the optimization of the CHARMM energy function, using the PSO algorithm. The dihedral angles of the side chain are provided using a three-layer artificial neural network, with a backward propagation algorithm (Ayan Datta et al., 2008).

Qian et al. addressed the first neural networks for secondary structure prediction. They worked in a network with 17 input groups that had 21 units per group, 40 hidden units and three output units. The use of neural networks has begun to evolve and different architectures have been studied aimed at a better precision of forecasting (Qian et al., 1988). **Chandonia et al.** also used the standard amino acid sequence as input for their neural network, and used the output of this prediction along with other information to predict the tertiary structure (Chandonia et al., 1995). Subsequently, HeiDelber introduced the prediction methodology of neural profile networks, which is considered the backbone of all current methods (Rost et al., 1996). **Gianluca Pollastri et al.** discussed more advanced neural networks. They used the two-way recurrent neuronal network. They also presented the second version of the SSpro program for forecasting the secondary structure (Gianluca Pollastri et al., 2002). **Glasgow et al.** have introduced a different technique in neural networks. They used the case-based reasoning technique to predict the secondary structure of the protein (Glasgow et al., 2006).

Sui et al. discussed the proposed method that hybrid SVM may improve the accuracy of secondary protein structure prediction. They proposed to combine the physiochemical properties of amino acid residues with position-specific scoring matrices that contain evolutionary information (Sui et al., 2011). **Li et al.** applied hierarchical structured SVM for protein folds prediction (Li et al., 2016). **Jian Guo et al.** proposed a high-performance method for predicting secondary protein structure based on the dual-layer carrier (SVM) machine and position-specific scoring matrices (PSSM) (Jian Guo et al., 2004).

Băutu et al. proposed a new discrete PSO variant for network-based protein folding models, such as the HP model. They present three lattice-based models and two folding coding, which are tested in different combinations on six proteins and found that this algorithm works very efficiently and produces excellent protein conformations (Băutu et al., 2010). **Guo et al.** proposed a new hybrid method to combine different heuristic strategies and biological problems in the optimization model to study the hydrophobic-hydrophilic model in a 3D square network (Guo et al., 2017). **Lin et al.** proposed Hybrid Genetic Algorithm and Particle Swarm Optimization for Protein 3D HP Model Folding Simulation (Lin et al., 2011). **Satpathy et al.** proposed bioinspired algorithms in solving Three-Dimensional PSP problems (Satpathy et al., 2017).

The Monte Carlo (MC) approach has also been widely used in the PSP. **Rout et al.** discussed the Boltzmann's law to sample possible conformations in the energy landscape and look for the optimal structure with minimum energy through iterative sampling, but the conformation tends to remain trapped in local minima at low temperatures, and so it often takes too long to reach convergence (Rout et al., 2013). Several variants have been developed to overcome this problem. **Liang et al.** proposed an evolutionary simulation approach of Monte Carlo (EMC) (Liang et al., 2001), which spends the research space by incorporating the genetic algorithm (GA) (Morris et al., 1998). Another variant called Monte Carlo entropy sampling, described by Lee (Lee et al., 1993) and then used in other studies incorporated a distribution based on entropy of protein conformations in the Monte Carlo simulation process (Kolinski et al., 1998; Scheraga et al., 2007).

5. Motivation

The protein structure prediction (PSP) is an NP-hard problem, which means no exact and efficient method can give the optimal solution to the problem in polynomial time (Chatterjee et al., 2019). PSP has many additional objectives to be optimized with several constraints to be considered, which leads to different problems of PSP that have been introduced in last three decades. Various approximation methods have already been proposed to solve the PSP by a number of researchers for decades. But several methods have limitations when the complexity of problem increases. To overcome these limitations, Nature Inspired Computing (NIC) based techniques may be used to solve PSP, which gives near optimal solutions to the problem in comparison to the other existing methods. The Swarm Intelligence (SI) is a subset of Nature Inspired Computing (NIC) emerged as a successful approach to solve large and complex optimization problems (like PSP) and also become an important research area to work on. Therefore motivated by both i.e. PSP's practical importance and considerable difficulty and advantageous features of SI techniques to solve complex optimization problems, the proposed research problems have been framed. The motivation is to implement NIC techniques to solve certain problems of the PSP in an efficient manner.

6. Objectives

The objective is to solve the problem of Protein Structure Prediction laid out in section 2 by using various Nature Inspired Computing (NIC) techniques and compare the results so obtained. It is proposed to:

1. Study how Nature Inspired techniques can be used to solve protein structure prediction (PSP) problem such as 2D and 3D structure prediction, protein folding, conformational search spaces etc.
2. Explore the possibility of parallel implementation of NI techniques to solve PSP problem.
3. Implementation of the newly developed techniques and algorithms in NI on the PSP under study.

4. Comparison of the newly developed techniques with existing techniques on standard benchmark problems.

7. Platform

MATLAB/Python will be used as the main tool to run simulations, and to compute and analyse the results of the newly developed algorithms.

8. References

Aarts, E., & Lenstra, J. K. (Eds.). (2003). Local Search in Combinatorial Optimization. Princeton University Press.

Al Kawam, Sen A, Datta A, Dickey N. Understanding the Bioinformatics Challenges of Integrating Genomics into Healthcare. IEEE Journal of Biomedical and Health Informatics. 2018; 22(5):1672-1683. doi: 10.1109/JBHI.2017.2778263.

Ali, A. F., & Hassanien, A. E. (2016). A survey of metaheuristics methods for bioinformatics applications. In Applications of Intelligent Optimization in Biology and Medicine, 23-46. Springer, Cham.

Angeleri, E., Apolloni, B., Falco, D. D., & Grandi, L. (1999). DNA fragment assembly using neural prediction techniques. International journal of neural systems, 9(06), 523-544.

Bai Li, Ya Li and Ligang Gong 2014. Protein secondary structure optimization using an improved artificial bee colony algorithm based on AB off-lattice model. Engineering Applications of Artificial Intelligence, vol. 27, 70–79.

Băutu, A., & Luchian, H. (2010, September). Protein structure prediction in lattice models with particle swarm optimization. In International Conference on Swarm Intelligence, 512-519. Springer, Berlin, Heidelberg.

Blanco, A., Pelta, D. A., & Verdegay, J. L. (2002). Applying a fuzzy sets- based heuristic to the protein structure prediction problem. International Journal of Intelligent Systems, 17(7), 629-643.

Chandonia, J. M., & Karplus, M. (1995). Neural networks for secondary structure and structural class predictions. Protein Science, 4(2), 275-285.

Chatterjee, S., & Shill, P. C. (2019, July). Protein Folding Optimization in a Hydrophobic-Polar Model for Predicting Tertiary Structure Using Fruit Fly Optimization Algorithm. In 2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT), 1-7. IEEE.

Chen, X., Lv, M., Zhao, L., & Zhang, X. (2011). An improved particle swarm optimization for protein folding prediction. International Journal of Information Engineering and Electronic Business, 3(1), 1.

Chen, Y., Pan, Y., Chen, J., Liu, W., & Chen, L. (2006, April). Partitioned optimization algorithms for multiple sequence alignment. In 20th International Conference on Advanced Information Networking and Applications-Volume 1 (AINA'06) (Vol. 2, pp. 5-pp). IEEE.

Cui, Z., & Zhang, Y. (2014). Swarm intelligence in bioinformatics: Methods and implementations for discovering patterns of multiple sequences. Journal of nanoscience and nanotechnology, 14(2), 1746-1757.

Das, S., Abraham, A., & Konar, A. (2008). Swarm intelligence algorithms in bioinformatics. In Computational Intelligence in Bioinformatics, 113-147. Springer, Berlin, Heidelberg.

Dill, K. A. (1985). Theory for the folding and stability of globular proteins. Biochemistry, 24(6), 1501-1509.

Dorigo, M., Caro, G. D., & Gambardella, L. M. (1999). Ant Algorithms for Discrete Optimization. Artificial Life, vol. 5 (2), 137-172.

Dubey, S. P., Balaji, S., Kini, N. G., & Kumar, M. S. (2016). A comparative study of various meta-heuristic algorithms for Ab initio protein structure prediction on 2D hydrophobic-polar model. In Proceedings of Fifth International Conference on Soft Computing for Problem Solving (pp. 387-399). Springer, Singapore.

Feng, D. F., & Doolittle, R. F. (1990). Progressive alignment and phylogenetic tree construction of protein sequences.

Gao, S., Song, S., Cheng, J., Todo, Y., & Zhou, M. (2018). Incorporation of solvent effect into multi-objective evolutionary algorithm for improved protein structure prediction. IEEE/ACM transactions on computational biology and bioinformatics, 15(4), 1365-1378.

Gianluca Pollastri, Darisz Przybylski, Burkhard Rost and Pierre Baldi, (2002, May) "Improving the Prediction of Protein Secondary Structure in Three and Eight Classes Using Recurrent Neural Networks and Profiles," PubMed, 228-235.

Glasgow, J., Kuo, T., & Davies, J. (2006). Protein structure from contact maps: A case-based reasoning approach. Information Systems Frontiers, 8(1), 29-36.

Glover, F., & Laguna, M. (1998). Tabu Search. Springer.

Guo, J., Chen, H., Sun, Z., & Lin, Y. (2004). A novel method for protein secondary structure prediction using dual- layer SVM and profiles. PROTEINS: Structure, Function, and Bioinformatics, 54(4), 738-743.

Guo, Y., Tao, F., Wu, Z., & Wang, Y. (2017). Hybrid method to solve HP model on 3D lattice and to probe protein stability upon amino acid mutations. BMC systems biology, 11(4), 93.

Hamam, Y., Al-Ani, T. 1996: Simulated annealing approach for Hidden Markov Models. In: Proc. 4th WG-7.6 Working Conf. on Optimization-Based Computer-Aided Modeling and Design, ESIEE, France

Hassanien, A. E., Al-Shammari, E. T., & Ghali, N. I. (2013). Computational intelligence techniques in bioinformatics. Computational biology and chemistry, 47, 37-47.

Indumathy, R., & Maheswari, S. U. (2012). Nature inspired algorithms to solve DNA fragment assembly problem: A survey. International Journal of Bioinformatics Research and Applications, 2(2), 45-50.

J. Setubal, and J. Meidanis, 2007. Introduction to Computational Molecular Biology, Boston: PWS Publishing Company.

Jong, K. D. (Ed.). (2006). Evolutionary computation: a unified approach. MIT Press.

Jozefowiez, N., Semet, F., & Talbi, E. G. (2002, September). Parallel and hybrid models for multi-objective optimization: Application to the vehicle routing problem. In International Conference on Parallel Problem Solving from Nature, 271-280. Springer, Berlin, Heidelberg.

Juang, W. S., & Su, S. F. (2008). Multiple sequence alignment using modified dynamic programming and particle swarm optimization. Journal of the Chinese institute of engineers, 31(4), 659-673.

Karaboga, D. (2005). An Idea Based on Honey Bee Swarm for Numerical Optimization. Turkey: Computer Engineering Department, Erciyes University.

Karaboga, D., & Akay, B. (2011). A Modified Artificial Bee Colony (Abc) Algorithm for Constrained Optimization Problems. Applied Soft Computing, 11 (3), 3021-3031.

Kennedy, J., & Eberhart, R. C. (1995). Particle Swarm Optimization. In Proceedings of the Fourth IEEE Intrenational Confrence on Neural Networks, 1942-1948. Perth, Australia: IEEE Service Center.

Kenneth, A., & Jong, D. (2006). Evolutionary Computation: a unified approach. MIT Press.

Khakzad, H., Karami, Y., & Arab, S. S. (2015). Accelerating protein structure prediction using particle swarm optimization on GPU. BioRxiv, 022434.

Khoury, G. A., Liwo, A., Khatib, F., Zhou, H., Chopra, G., Bacardit, J., ... & Krupa, P. (2014). WeFold: a coopetition for protein structure prediction. Proteins: Structure, Function, and Bioinformatics, 82(9), 1850-1868.

Kitagawa, J., & Iba, H. (2003). Identifying metabolic pathways and gene regulation networks with evolutionary algorithms. In Evolutionary computation in bioinformatics (pp. 255-278). Morgan Kaufmann.

Kiyoshi Asai, Satoru Hayamizu, and KI Handa. "Prediction of Protein Secondary Structure by the Hidden Markov Model." In: 9 (May 1993).

Kolinski, A., Galazka, W. and Skolnick, J. Monte Carlo studies of the thermodynamics and kinetics of reduced protein models: Application to small helical, β, and α/β proteins. The Journal of chemical physics 1998;108(6):2608- 2617.

Krogh, A., Brown, M., Mian, I. S., Sjölander, K., & Haussler, D. (1994). Hidden Markov models in computational biology: Applications to protein modeling. Journal of molecular biology, 235(5), 1501-1531.

Laarhoven, P. J., & Aarts, E. H. (Eds.). (1987). Simulated annealing: theory and applications. Kluwer Academic Publishers.

Lalwani, S., Kumar, R., & Gupta, N. (2015). A novel two-level particle swarm optimization approach for efficient multiple sequence alignment. Memetic Computing, 7(2), 119-133.

Lamiche Chaabane,(2018). A Hybrid Solver for Protein Multiple Sequence Alignment Problem, Journal of Bioinformatics and Computational Biology.

Lau, K. F., & Dill, K. A. (1989). A lattice statistical mechanics model of the conformational and sequence spaces of proteins. Macromolecules, 22(10), 3986-3997.

Lee, J. New Monte Carlo algorithm: Entropic sampling. Phys Rev Lett 1993;71(2):211-214.

Lesk, A. (2019). Introduction to bioinformatics. Oxford university press.

Li, B., Li, Y., & Gong, L. (2014). Protein secondary structure optimization using an improved artificial bee colony algorithm based on AB off-lattice model. Engineering Applications of Artificial Intelligence, 27, 70-79.

Li, D., Ju, Y., & Zou, Q. (2016). Protein folds prediction with hierarchical structured SVM. Current Proteomics, 13(2), 79-85.

Liang, F. and Wong, W.H. Evolutionary Monte Carlo for protein folding simulations. Journal of Chemical Physics 2001;115(7):3374-3380.

Lin, C. J., & Su, S. C. (2011). Protein 3D HP Model Folding Simulation Using a Hybrid of Genetic Algorithm and Particle Swarm Optimization. International Journal of Fuzzy Systems, 13(2).

Lipman DJ, Altschul SF and Kececioglu JD. (1989). A tool for multiple sequence alignment. Proc. Natl. Acad. Sci. USA, 86:4412–4415.

Liu, J., Wang, L., He, L., Shi, F (2005). In: Analysis of Toy Model for Protein Folding Based on Particle Swarm Optimization Algorithm. Volume 3612. Springer Berlin/Heidelberg 636-645.

Long, H. X., Wu, S. L., & Lv, Y. (2014). Protein structure prediction based on profile HMM and QPSO. In Advanced Materials Research (Vol. 1004, pp. 853-856). Trans Tech Publications Ltd.

Mathur, M. (2018). Bioinformatics challenges: A review. Bioinformatics, 3(6).

Meksangsouy, P., & Chaiyaratana, N. (2003, December). DNA fragment assembly using an ant colony system algorithm. In The 2003 Congress on Evolutionary Computation, 2003. CEC'03. Vol. 3, 1756-1763. IEEE.

Morris, G.M., et al. (1998). Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. Journal of computational chemistry; 19(14):1639-1662.

Moult, J., Fidelis, K., Kryshtafovych, A., Schwede, T., & Tramontano, A. (2016). Critical assessment of methods of protein structure prediction: Progress and new directions in round XI. Proteins: Structure, Function, and Bioinformatics, 84, 4-14.

Nasser, S., Vert, G.L., Nicolescu, M., Murray, A. 2007: Multiple Sequence Alignment using Fuzzy Logic. In: Proc. IEEE Symposium on Computational Intelligence and Bioinformatics and Computational Biology, 304–311.

Osman, I. H., & Laporte, G. (1996). Metaheuristics: A bibliography. Ann. Oper. Res., 63, 513-623.

Parpinelli, R. S., Benitiez, C. M., Cordeiro, J., & Lopes, H. S. (2014). Performance Analysis of Swarm Intelligence Algorithms for the 3D-AB off-lattice Protein Folding Problem. Multiple-Valued Logic and Soft Computing, 22(3), 267-286.

Pelta, D., Krasnogor, N., Bousono-Calzon, C., Verdegay, J. L., Hirst, J., & Burke, E. (2005).A fuzzy sets based generalization of contact maps for the overlap of protein structures. Fuzzy Sets and Systems, 152(1), 103-123.

Pevsner, J. (2015). Bioinformatics and functional genomics. John Wiley & Sons.

Pfeiffenberger, E., & Bates, P. A. (2018). Predicting improved protein conformations with a temporal deep recurrent neural network. PloS one, 13(9).

Qian, N., & Sejnowski, T. J. (1988). Predicting the secondary structure of globular proteins using neural network models. Journal of molecular biology, 202(4), 865-884.

Rahman, A., & Zomaya, A. Y. (2005). An overview of protein-folding techniques: issues and perspectives. International Journal of Bioinformatics Research and Applications, 1(1), 121-143.

Rajakumar, R., Dhavachelvan, P., & Vengattaraman, T. (2016, October). A survey on nature inspired meta-heuristic algorithms with its domain specifications. In 2016 International Conference on Communication and Electronics Systems (ICCES), 1-6. IEEE.

Rasmussen, T. K., & Krink, T. (2003). Improved Hidden Markov Model training for multiple sequence alignment by a particle swarm optimization—evolutionary algorithm hybrid. Biosystems, 72(1-2), 5-17.

Rayapudi, S. R. (2011). An Intelligent Water Drop Algorithm for Solving Economic Load Dispatch Problem. International Journal of Electrical and Electronics Engineering, 5 (1), 43-49.

Riaz T, Wang Y, Li KB. 2004. Multiple sequence alignment using tabu search, Proceedings of 2nd Asia-Pacific Bioinformatics Conference (APBC), New Zealand, 223–232.

Richards, F. M. (1977). Areas, volumes, packing, and protein structure. Annual review of biophysics and bioengineering, 6(1), 151-176.

Rost, B. (1996). [31] PHD: Predicting one-dimensional protein structure by profile-based neural networks. In Methods in enzymology, Vol. 266, 525-539. Academic Press.

Rout, S. B., Dehury, S., & Mishra, B. S. P. (2013). Protein Structure Prediction using Genetic Algorithm.

Rubio-Largo, Á., Vega-Rodríguez, M. A., & González-Álvarez, D. L. (2016). Hybrid multiobjective artificial bee colony for multiple sequence alignment. Applied Soft Computing, 41, 157-168.

Samudrala, R., Xia, Y., Levitt, M., & Huang, E. S. (1999). A combined approach for ab initio construction of low-resolution protein tertiary structures from sequence. In Biocomputing'99, 505-516.

Satpathy, R. (2017). Bioinspired Algorithms in Solving Three-Dimensional Protein Structure Prediction Problems. In Bio-Inspired Computing for Information Retrieval Applications, 316-337. IGI Global.

Scheraga, H.A. and Hao, M.H. (2007). Entropy Sampling Monte Carlo for Polypeptides and Proteins.

Setubal J and Meidanis J. (1999) Introduction to Computational Molecular Biology, International Thomson Publishing, 20 park plaza, Boston, MA 02116.

Shah-Hosseini, H. (2007). Problem Solving by Intelligent Water Drops. Proc. IEEE Congress on Ecolutionary Computation, (pp. 3226-3231). Singapore.

Shah-Hosseini, H. (2009). The intelligent water drops algorithm: a nature-inspired swarm-based optimization algorithm. International Journal of Bio-Inspired Computation, 1(1-2), 71-79.

Stolcke, A., & Omohundro, S. (1993). Hidden Markov model induction by Bayesian model merging. In Advances in neural information processing systems, 11-18.

Sui, H., Qu, W., Yan, B., & Wang, L. (2011). Improved protein secondary structure prediction using an intelligent HSVM method with a new encoding scheme. International Journal of Advancements in Computing Technology, 3(3), 239-250.

Tan, K. C., Chew, Y. H., & Lee, L. H. (2006). A Hybrid Multi-Objective Evolutionary Algorithm for Solving Vehicle Routing Problem with Time Window. Computational Optimization and Applications, 34 (1), 115-151.

Verma, R. S. (2012, March). DSAPSO: DNA sequence assembly using continuous particle swarm optimization with smallest position value rule. In 2012 1st international conference on recent advances in information technology (RAIT) (pp. 410-415). IEEE.

Verma, R. S., Singh, V., & Kumar, S. (2011). Dna sequence assembly using particle swarm optimization. International Journal of Computer Applications, 28(10), 33-38.

Wang, Y., Mao, H., & Yi, Z. (2017). Protein secondary structure prediction by using deep learning method. Knowledge-Based Systems, 118, 115-123.

Wetcharaporn, W., Chaiyaratana, N., & Tongsima, S. (2006, June). DNA fragment assembly by ant colony and nearest neighbour heuristics. In International Conference on Artificial Intelligence and Soft Computing (pp. 1008-1017). Springer, Berlin, Heidelberg.

Zhang, G.-Z., Huang, D.-S. 2004: Aligning multiple protein sequence by an improved genetic algorithm. In: Proc. IEEE Intl. Joint Conf. on Neural Networks, 1179– 1183.

Zhang, Y., & Wu, L. (2012). Artificial bee colony for two-dimensional protein folding. Advances in Electrical Engineering Systems, 1(1), 19-23.

Zhou, C., Hou, C., Wei, X., & Zhang, Q. (2014). Improved hybrid optimization algorithm for 3D protein structure prediction. Journal of molecular modeling, 20(7), 2289.

Zhou, C., Hou, C., Zhang, Q., & Wei, X. (2013). Enhanced hybrid search algorithm for protein structure prediction using the 3D-HP lattice model. Journal of molecular modeling, 19(9), 3883-3891.

Zhu, H., Pu, C., Lin, X., Gu, J., Zhang, S., Su, M. (2009): Protein structure prediction with EPSO in Toy model. In: Proc. of ICINIS '09, Washington, DC, USA, IEEE Computer Society 673-676.